Adaptive Gaussian Predictive Process Approximation

Surya T Tokdar
Duke University

Abstract

We address the issue of knots selection for Gaussian predictive process methodology. Predictive process approximation provides an effective solution to the cubic order computational complexity of Gaussian process models. This approximation crucially depends on a set of points, called knots, at which the original process is retained, while the rest is approximated via a deterministic extrapolation. Knots should be few in number to keep the computational complexity low, but provide a good coverage of the process domain to limit approximation error. We present theoretical calculations to show that coverage must be judged by the canonical metric of the Gaussian process. This necessitates having in place a knots selection algorithm that automatically adapts to the changes in the canonical metric affected by changes in the parameter values controlling the Gaussian process covariance function. We present an algorithm toward this by employing an incomplete Cholesky factorization with pivoting and dynamic stopping. Although these concepts already exist in the literature, our contribution lies in unifying them into a fast algorithm and in using computable error bounds to finesse implementation of the predictive process approximation. The resulting adaptive predictive process offers a substantial automatization of Guassian process model fitting, especially for Bayesian applications where thousands of values of the covariance parameters are to be explored.

Keywords: Gaussian predictive process, Knots selection, Cholesky factorization, Pivoting, Bayesian model fitting, Markov chain sampling.

1 Introduction

Bayesian nonparametric methodology is driven by construction of prior distributions on function spaces. Toward this, Gaussian process distributions have proved extremely useful due to their mathematical and computational tractability and ability to incorporate a wide range of smoothness assumptions. Gaussian process models have been widely used in spatio-temporal modeling (Handcock and Stein, 1993; Kim et al., 2005; Banerjee et al., 2008), computer emulation (Kennedy and O'Hagan, 2001; Oakley and OHagan, 2002), non-parametric regression and classification (Neal, 1998; Csató et al., 2000; Rasmussen and Williams, 2006; Short et al., 2007), density and quantile regression (Tokdar et al., 2010; Tokdar and Kadane, 2011), functional data analysis (Shi and Wang, 2008; Petrone et al., 2009), image analysis (Sudderth and Jordan, 2009), etc. Rasmussen and Williams (2006) give a thorough overview of likelihood based exploration of Gaussian process models, including Bayesian treatments. For theoretical details on common Bayesian models based on Gaussian processes, see Tokdar and Ghosh (2007), Choi and Schervish (2007), Ghosal and Roy (2006), van der Vaart and van Zanten (2008, 2009) and the references therein.

For our purpose, a Gaussian process can be viewed as a random, real valued function $\omega = (\omega(t), t \in T)$ on a compact Euclidean domain T, such that for any finitely many points $t_1, \cdots, t_k \in T$ the random vector $(\omega(t_1), \cdots, \omega(t_k))$ is a k-dimensional Gaussian vector. A Gaussian process ω is completely characterized by its mean and covariance functions $\mu(t) = \mathrm{E}[\omega(t)]$ and $\psi(s,t) = \mathrm{Cov}[\omega(s), \omega(t)]$. For a Gaussian process model, where a function valued parameter ω is assigned a Gaussian process prior distributions, the data likelihood typically involves ω through a vector $W = (\omega(s_1), \cdots, \omega(s_N))$ of ω values at a finite set of points $s_1, \cdots, s_N \in T$. These points could possibly depend on other model parameters. The fact that W is a Gaussian vector makes computation conceptually straightforward.

However, a well known bottleneck in implementing Gaussian process models is the $O(N^3)$ complexity of inverting or factorizing the $N \times N$ covariance matrix of W. Various reduced rank approximations to covariance matrices have been proposed to overcome this problem (Smola and Bartlett, 2001; Seeger et al., 2003; Schwaighofer and Tresp, 2003; Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006), mostly reported in the machine learning literature. Among these, a special method of approximation, known as predictive process approximation (Tokdar, 2007; Banerjee et al., 2008), has been independently discovered and successfully used in the Bayesian literature. The appeal of this method lies in a stochastic process representation of the approximation that obtains from tracking ω at a small number of points, called knots, and extrapolating the rest by using properties of Gaussian process laws (Section 2.1).

For predictive process approximations, choosing the number and locations of the knots remains a difficult problem. This problem is only going to escalate as more complex Gaussian process models are used in hierarchical Bayesian modeling, with rich parametric and non-parametric formulations of the Gaussian process covariance function becoming commonplace. To understand this difficulty, we first lay out (Section 2.2) the basic probability theory behind the approximation accuracy of the predictive process and demonstrate how the choice of knots determines an accuracy bound. The key concept here is that the knots must provide a good coverage of the domain T of the Gaussian process. While this is intuitive, what needs emphasis is that the geometry of T is to be viewed through the topology induced by the Gaussian process canonical metric $\rho(s,t) = [\mathrm{E}\{\omega(s) - \omega(t)\}^2]^{1/2}$, which could be quite different from the Euclidean geometry of T.

This theory helps understand (Section 2.3) why existing approaches of choosing knots, based on space filling design concepts (Zhu and Stein, 2005; Zimmerman, 2006; Finley et al., 2009) or model extensions where knots are learned from data as additional model parameters (Snelson and Ghahramani, 2006; Tokdar, 2007; Guhaniyogi et al., 2010), are likely to offer poor approximation and face severe computational difficulties. A fundamental weakness of these approaches is their inability to automatically adapt to the changes in the geometry of T caused by changes in the values of the covariance parameters.

In Section 3 we present a simple extension of the predictive process approximation that enables it to automatically adapt to the geometry of the Gaussian process covariance function. This extension, called *adaptive predictive process approximation*, works with an equivalent representation of the predictive process through reduced rank Cholesky factorization (Section 3.1) and adds to it two adaptive features, *pivoting* and *dynamic stopping*. Pivoting determines the order in which knots are selected from an initial set of candidate points while dynamic stopping determines how many knots to select. The resulting approximation meets a pre-

specified accuracy bound (Section 3.2).

The connection between predictive process approximation and reduced rank Cholesky factorization is well known (Quiñonero-Candela and Rasmussen, 2005) and pivoting has been recently investigated in this context from the point of view of stable computation (Foster et al., 2009). The novelty of our work lies in unifying these ideas to define an adaptive version of the predictive process and in proposing accuracy bounds as the driving force in finessing the implementation of such approximation techniques. The end product is a substantial automatization of fitting Gaussian process models that can broaden up the scope of such models without the additional burden of having to model or learn the knots. This is illustrated with two examples in Section 4.

2 Predictive process approximation

2.1 Definition

Fix a set of points, referred to as *knots* hereafter, $\{t_1, t_2, \cdots, t_m\} \subset T$ and write $\omega(t) = \nu(t) + \xi(t)$, where,

$$\nu(t) = E\{\omega(t) \mid \omega(t_1), \cdots, \omega(t_m)\}$$

and $\xi(t) = \omega(t) - E\{\omega(t)|\omega(t_1), \cdots, \omega(t_m)\}$. By the properties of Gaussian process laws, ν and ξ are independent Gaussian processes. The process ν, called a Gaussian predictive process, has rank m, because it can be written as $\nu(t) = \sum_{i=1}^{m} A_i \psi(t_i, t)$ with the coefficient vector $A = (A_1, \cdots, A_m)$ being a Gaussian vector. By replacing ω with ν in the statistical model, one now deals with the covariance matrix of $V = (\nu(s_1), \cdots, \nu(s_N))$, which, due to the rank-m property of ν, can be factorized in $O(Nm^2)$ time.

2.2 Accuracy bound

Replacing ω with ν can be justified as follows. Let $\delta = \sup_{t \in T} \min_{1 \le i \le m} \rho(t, t_i)$ denote the mesh size of the knots, where $\rho(t, s) = [E\{\omega(t) - \omega(s)\}^2]^{1/2}$ is the canonical metric on T induced by ω (Adler, 1990, page 2). For a smooth $\psi(t, s)$, δ can be made arbitrarily small by packing T with sufficiently many, well placed knots. But, as δ tends to 0, so does $\kappa^2 = \sup_{t \in T} \mathrm{Var}\{\xi(t)\}$. This is because for any $t \in T$, and any $i \in \{1, \cdots, m\}$, by the independence of ν and ξ,

$$\begin{aligned}
\mathrm{Var}\{\xi(t)\} &= \mathrm{Var}\{\omega(t)\} - \mathrm{Var}\{\nu(t)\} \\
&= \mathrm{E}[\mathrm{Var}\{\omega(t)|\omega(t_1), \cdots, \omega(t_m)\}] \\
&= \mathrm{E}[\mathrm{Var}\{\omega(t) - \omega(t_i)|\omega(t_1), \cdots, \omega(t_m)\}] \\
&\le \mathrm{Var}\{\omega(t) - \omega(t_i)\} = \rho^2(t, t_i),
\end{aligned}$$

and hence $\kappa \le \delta$. That κ can be made arbitrarily small is good news, because it plays a key role in providing probabilistic bounds on the residual process ξ.

Theorem 2.1. *Let ω be a zero mean Gaussian process on a compact subset $T \subset R^p$. Let ν be a finite rank predictive process approximation of ω with residual process $\xi = \omega - \nu$.*

(i) If $T \subset [a,b]^p$ and there is a finite constant $c > 0$ such that $\mathrm{Var}\{\omega(s)-\omega(t)\} \leq c^2\|s-t\|^2$, $s,t \in T$ then

$$P\left(\sup_{t\in T}|\xi(t)| > \epsilon\right) \leq 3\exp\left(-\frac{\epsilon^2}{B^2\kappa}\right), \quad \forall \epsilon > 0 \tag{1}$$

with $B = 27\sqrt{2pc(b-a)}$ and $\kappa^2 = \sup_{t\in T}\mathrm{Var}\{\xi(t)\}$.

(ii) For any finite subset $S \subset T$

$$P\left(\sup_{t\in S}|\xi(t)| > \epsilon\right) \leq 3\exp\left\{-\frac{\epsilon^2}{9\kappa_S^2(2+\log|S|)}\right\}, \quad \forall \epsilon > 0 \tag{2}$$

where $|S|$ denotes the size of S and $\kappa_S^2 = \sup_{t\in S}\mathrm{Var}\{\xi(t)\}$.

A proof is given in Appendix A. Note that the constant B does not depend on the number or locations of the knots, it only depends on the dimensionality and size of T as well as smoothness properties of the covariance function ω. It is possible to replace κ in (1) with $\kappa^{2(1-\eta)}$ for any arbitrary $\eta \in (0,1)$, but with a different constant B. While (1) provides an accuracy bound over the entire domain T, the bound in (2) over a finite subset maybe of more practical value.

For Gaussian process regression models with additive Gaussian noise, a common modification (Finley et al., 2009) of predictive process approximation is to replace ω with the process $\tilde{\nu} = \nu + \xi^*$ where ξ^* is a zero mean Gaussian process, independent of ν and ξ, satisfying,

$$\mathrm{Cov}\{\xi^*(t),\xi^*(s)\} = \begin{cases} \mathrm{Cov}\{\xi(t),\xi(s)\} = \mathrm{Var}\xi(t) & \text{if } t = s \\ 0 & \text{if } t \neq s. \end{cases}$$

The addition of ξ^* gives $\tilde{\nu}$ the same pointwise mean and variance as those of ω, without adding to the computational cost. The residual process is now given by $\tilde{\xi} = \omega - \tilde{\nu} = \xi - \xi^*$ whose variance equals $2\mathrm{Var}\{\xi(t)\}$ because of independence between ξ and ξ^*. Because ξ^* is almost surely discontinuous, the bound in (1) does not apply to $\tilde{\xi}$. But (2) continues to hold with κ_S^2 replaced by $\tilde{\kappa}_S^2 = 2\kappa_S^2$.

2.3 Need for adaptation

For predictive process approximations, choosing the number and the locations of the knots remains a difficult problem. Ideally this choice should adapt to the canonical metric ρ, so that a small δ obtains with as few knots as possible. However, it's not a single ρ that we need to adapt to. In modern Gaussian process applications, the covariance function ψ and consequently the canonical metric depend on additional model parameters. A typical example is ω of the form

$$\omega(t) = \omega_0(t) + x_1(t)\omega_1(t) + \ + x_p(t)\omega_p(t)$$

where $x_j(t)$'s are known, fixed functions and ω_j's are independent mean zero Gaussian processes with covariances $\psi_j(t,s) = \tau_j^2\exp(-\beta_j^2\|s-t\|^2)$ with $\theta = (\tau_0,\tau_1,\cdots,\tau_p,\beta_0,\beta_1,\cdots,\beta_p)$ serving as a model parameter. Because a likelihood based model fitting will loop through hundreds or even thousands of values of θ, it is important to have a low-cost algorithm to choose the knots that automatically adapts to the geometry of any arbitrary canonical metric.

Such an adaptive feature is lacking from existing knots selection approaches which primarily treat knots as additional model parameters. The knots are then learned along with other model parameters, either via optimization (Snelson and Ghahramani, 2006) or by Markov chain sampling (Tokdar, 2007; Guhaniyogi et al., 2010). Another popular approach is to work with a fixed set of knots based on space-filling design concepts (Zhu and Stein, 2005; Zimmerman, 2006; Finley et al., 2009). Among these, the proposal in Finley et al. (2009) overlaps with our proposal. But while we pursue a low-cost adaptation at every value of θ at which likelihood evaluation is needed, Finley et al. (2009) consider one fixed set of knots adapted to a representative value of θ. Their knot selection algorithm has $O(N^2m)$ computing time, which makes it infeasible to run within an optimization or a Markov chain sampler loop.

3 Adaptative predictive process

3.1 Predictive process approximation via Cholesky factorization

Let $\omega = (\omega(t), t \in T)$ be a zero-mean Gaussian process with covariance $\psi(s,t)$. Suppose the finite set $S = \{s_1, s_2, \cdots, s_N\} \subset T$ contains all points in T where ω needs to be evaluated for the purpose of model fitting. Let $\Psi = ((\psi_{ij}))$ denote the $N \times N$ covariance matrix of the Gaussian vector $W = (\omega(s_1), \cdots, \omega(s_N))$. A Cholesky factor Λ of Ψ, with Λ being a $N \times N$ upper triangular matrix with non-negative diagonal elements and satisfying $\Psi = \Lambda'\Lambda$, obtains from the following recursive calculations

$$\lambda_{ii} = \sqrt{\psi_{ii} - \sum_{\ell<i} \lambda_{\ell i}^2}, \;\; \lambda_{ij} = \frac{\psi_{ij} - \sum_{\ell<i} \psi_{\ell j}\psi_{\ell i}}{\lambda_{ii}}, \;\; i = 1, \cdots, N, j = i+1, \cdots, N \qquad (3)$$

with λ_{ij}, $j < i$ set to zero. This gives a row-by-row construction of Λ and requires $C_M N i^2$ computation time for constructing the first i rows for some machine dependent constant C_M.

For an $m \in \{1, \cdots, N\}$, an approximation $\hat{\Lambda} = ((\hat{\lambda}_{ij}))$ to Λ obtains in $O(Nm^2)$ time by an incomplete application of the above recursion. The first m rows of $\hat{\Lambda}$ are constructed in $C_M N m^2$ time through (3):

$$\hat{\lambda}_{ii} = \sqrt{\psi_{ii} - \sum_{\ell<i} \{\hat{\lambda}_{\ell i}\}^2}, \;\; \hat{\lambda}_{ij} = \frac{\psi_{ij} - \sum_{\ell<i} \psi_{\ell j}\psi_{\ell i}}{\hat{\lambda}_{ii}}, \;\; i = 1, \cdots, m, j = i+1, \cdots, N. \qquad (4)$$

For the remaining rows, only the diagonal elements are computed in $C_M(N-m)$ time as in the first part of (3), with the off diagonals set to zero:

$$\hat{\lambda}_{ii} = \sqrt{\psi_{ii} - \sum_{\ell \le m} \{\hat{\lambda}_{\ell i}\}^2}, \;\; \hat{\lambda}_{ij} = 0, \;\; i = m+1, \cdots, N, j = i+1, \cdots, N. \qquad (5)$$

The lower triangular elements $\hat{\lambda}_{ij}$, $j < i$, are all set to 0. The resulting $\hat{\Lambda}$ is upper triangular with non-negative diagonals and equals the Cholesky factor of the covariance matrix $\hat{\Psi}$ of $V = (\nu(s_1), \cdots, \nu(s_N))$ where $\nu = (\nu(t), t \in T)$ is the Gaussian predictive process approximation of ω based on knots $s_1, \cdots, s_m$. The resulting residual process $\xi = \omega - \nu$ satisfies $\mathrm{Var}\{\xi(s_i)\} = 0$, $1 \le i \le m$ and $\mathrm{Var}\{\xi(s_i)\} = \{\hat{\lambda}_{ii}\}^2$, $i = m+1, \cdots, N$, and hence $\kappa_S^2 := \max_{s \in S} \mathrm{Var}\{\xi(s)\} =$

$\max_{i>m} \hat{\lambda}_{ii}^2$. From (2), κ_S^2 controls error bounds $P(\max_{s\in S} |\omega(s) - \nu(s)| > \epsilon)$ over the set of interest S.

Therefore, the above incomplete Cholesky factorization produces a Gaussian predictive process approximation, with readily available error bounds, provided we are happy to choose the knots from the set S. The restriction to S appears reasonable for most applications with the additional burden on the modeler to identify S carefully. For example, in a Gaussian process regression model with additive noise, it is sufficient to take S to be the training set of covariate values, if only posterior predictive mean and variances are needed at test cases. But if posterior predictive covariance between two test cases, or a test and a training case is required, then S should include these test cases as well.

3.2 Pivoting and dynamic stopping

Our quest of an adaptive version of ν stays within this restriction, but employs a dynamic choice of the stopping time m and the order in which the elements of S are processed. To decide whether the current stopping time is acceptable, we check the current κ_S against a given tolerance κ_{tol}. If κ_S exceeds the tolerance, we increment m to $m+1$, and repeat (4) only for i equal to the new value of m, followed by (5), producing an update of κ_S. The top row elements $\hat{\lambda}_{ij}$, $i < m$ need no changes.

The increment of m and the subsequent alterations to $\hat{\Lambda}$ clearly reduce κ_S as the tailing $\hat{\lambda}_{ii}$'s in (5) are reduced. This reduction can be expected to improve if after incrementing m and before proceeding with the new calculations, one swaps the current m-th and k-th rows of Ψ where $i = k$ gives the maximum of the tailing $\hat{\lambda}_{ii}$ values, $i = m, \cdots, N$. A sequence of such swaps, from start to the terminating m, gives a greedy, dynamic approximation to finding the optimal ordering of the elements of S that gives a $\kappa_S \leq \kappa_{\mathsf{tol}}$ with a minimum stopping time m.

The dynamic swapping is a common feature, known as *pivoting*, of all leading software packages for Cholesky factorization. If run until $m = N$, pivoting produces a permutation $\pi = (\pi_1, \cdots, \pi_N)$ of $(1, \cdots, N)$ and an upper triangular matrix Λ with non-negative diagonals such that $P_\pi \Psi P_\pi' = \Lambda'\Lambda$ where P_π is the $N \times N$ permutation matrix associated with π. Our proposal above simply adds to this pivoted Cholesky factorization a dynamic, tolerance based stopping. The resulting $\hat{\Lambda}$ gives the Cholesky factor of the covariance matrix of $V = (\nu(s_1), \cdots, \nu(s_N))$ where ν is the Gaussian predictive process associated with the knots $s_{\pi_1}, \cdots, s_{\pi_m}$. The Gaussian predictive process ν comes with the error bound (2) with κ_{tol} replacing κ. The additional computing time needed for pivoting is only $O(Nm)$, a small fraction of the computing time $O(Nm^2)$ needed to get the elements $\hat{\Lambda}$ if π was precomputed.

Algorithm 1 gives a pseudo code for performing the incomplete Cholesky factorization with an additional improvisation. The user specified tolerance κ_{tol} is taken to be a relative tolerance level instead of an absolute one. The absolute tolerance is fixed as κ_{tol} times the maximum of $\mathrm{Var}\{\omega(s_i)\}^{1/2}$, $i = 1, \cdots, N$. This makes sense for Gaussian process approximation as the maximum variance of the process can be viewed as a scaling parameter.

3.3 Geometric Illustration

We illustrate the adaptive choice of knots on the Bartlett experimental forest dataset (Finley et al., 2009). This dataset contain $n = 437$ well identified forest locations $s_1, \cdots, s_n$, measured

Algorithm 1 Pivoted, incomplete Cholesky factorization with dynamic stopping.

Require: A covariance function $\psi(\cdot,\cdot)$, positive integers N and $m_{\max}$ and tolerance $\kappa_{\mathsf{tol}} > 0$.

$R \leftarrow 0_{m_{\max} \times m_{\max}}$
$\pi_1 \leftarrow 1, \pi_2 \leftarrow 2, \cdots, \pi_N \leftarrow N$
$k \leftarrow 1$
$d_{\max} \leftarrow \max_{1 \le l \le N} \psi(s_{\pi_l}, s_{\pi_l})$
$l_{\max} \leftarrow \arg\max_{1 \le l \le N} \psi(s_{\pi_l}, s_{\pi_l})$
$\kappa_{\mathsf{tol}} \leftarrow \sqrt{d_{\max}}\kappa_{\mathsf{tol}}$
while $d_{\max} > \kappa_{\mathsf{tol}}^2$ **do**
  swap π_k and $\pi(l_{\max})$
  $r_{kk} \leftarrow d_{\max}^{1/2}$
  for $j = k+1$ to m **do**
    $r_{kj} \leftarrow \{\psi(s_{\pi_k}, s_{\pi_j}) - \sum_{l<k} r_{lk} r_{lj}\}/r_{kk}$
  end for
  $k \leftarrow k+1$
  $d_{\max} \leftarrow \max_{k \le l \le N} \psi(s_{\pi_l}, s_{\pi_l}) - \sum_{l<k} r_{lk}^2$
  $l_{\max} \leftarrow \arg\max_{k \le l \le N} \psi(s_{\pi_l}, s_{\pi_l}) - \sum_{l<k} r_{lk}^2$
end while
$m \leftarrow k$
for $k = m+1$ to N **do**
  $r_{kk} \leftarrow \{\psi(s_{\pi_k}, s_{\pi_k}) - \sum_{l \le m} r_{lm}^2\}^{1/2}$
end for
return Factor matrix R, pivot π and rank m

as eastings (on the horizontal axis) and northings (vertical axis) from a reference point (Figure 1). For illustration purposes, we consider only a hypothetical Gaussian process model where a surface $\omega(s)$ over the forest area is modeled as a zero-mean Gaussian process with a square-exponential covariance function

$$\psi(s,t) = x(s)x(t)\exp(-\beta\|Q(s-t)\|^2), \tag{6}$$

for some constant $\beta > 0$, an orthogonal projection matrix Q and some fixed function $x(t)$ that relates to the slope of the forest landscape at location t. The variance of $\omega(t)$ is $x(t)^2$ and the correlation between $\omega(s)$ and $\omega(t)$ depends on the Euclidean distance between the Q-projections of these two location vectors. The parameter $\beta > 0$ encodes the spatial range of the covariance, i.e., it controls how rapidly $\psi(s,t)$ decays with the distance between s and t. We demonstrate how the choice of knots according to Algorithm 1 adapts to variations in each of β, $x(t)$ and Q.

Figure 1(a) shows nine choices of (6) that vary in Q and β, while $x(t)$ and κ_{tol} are held fixed at $x(t) \equiv 1$ and $\kappa^2_{\mathsf{tol}} = 10^{-4}$. The top row has $\beta = 10^{-3}$, the middle row has $\beta = 5\times10^{-4}$ and the bottom row has $\beta = 10^{-4}$. The left column has Q equal to the identity matrix, the middle column has Q that projects along the horizontal axis and the right column has Q that projects along the vertical axis (indicated by arrows). It is clear that the algorithm picks fewer knots as the spatial range decreases. A smaller spatial range gives a flatter topology in the canonical metric, consequently fewer points are required to capture the variation in the random surface $\omega(t)$. It is also clear that the algorithm adapts to the directional element of this topology. It lines up the knots along the horizontal axis when Q is the horizontal projection and lines up the knots along the vertical axis when Q projects along that axis.

Figure 1(b) shows nine other choices of (6) that vary in Q and $x(t)$, while β and κ_{tol} are held fixed at $\beta = 10^{-3}$ and $\kappa^2_{\mathsf{tol}} = 10^{-4}$. Variation in Q is as in Figure 1(a). We take $x(t) = 1 - F(c\cdot\texttt{slope}(t)|a,b)$, where $F(x|a,b)$ denotes the gamma distribution function with shape parameter a and rate parameter b, and $\texttt{slope}(t)$ is the slope of the landscape at point t. We fix a and b so that the mean and variance of the gamma distribution match the mean and variance of the recorded slope values at the 437 locations. The top row of Figure 1(b) has $c = 0$, the middle row has $c = 1$ and the bottom row has $c = 6$. Larger values of c make $x(t)$ closer to zero at regions with a high slope while $x(t)$ always equals 1 at regions that are flat (shown in lighter color, mostly along a narrow valley running from south-east to north-west). With a larger c, most of the variation in $\omega(t)$ is confined to locations t with a flat slope. It is clear that our algorithm adapts to this feature by picking knots from such areas.

Figure 1(c) shows (6) with $\beta = 10^{-4}$, $x(t) \equiv 1$, $Q =$ the identity matrix, but with four choices of $\kappa^2_{\mathsf{tol}} = 10^{-1}, 10^{-2}, 10^{-4}$ and 10^{-8} (clockwise from top left). For smaller tolerance levels, more knots are picked to meet a tighter accuracy condition. A good coverage of the entire region is maintained throughout, but additional knots are chosen to give a denser representation. Note the higher concentration of knots at the boundary than the interior. This is a consequence of the greedy nature of the algorithm as it tries to pick the next knot as the point that is least correlated with the ones already selected. Although a better algorithm could correct for such a boundary bias, the linear additional computing cost of the greedy search offers a highly attractive trade-off against a few extra knots. Figure 1(d) indicate that the number of knots m required to meet the tolerance criterion grows at a logarithmic rate

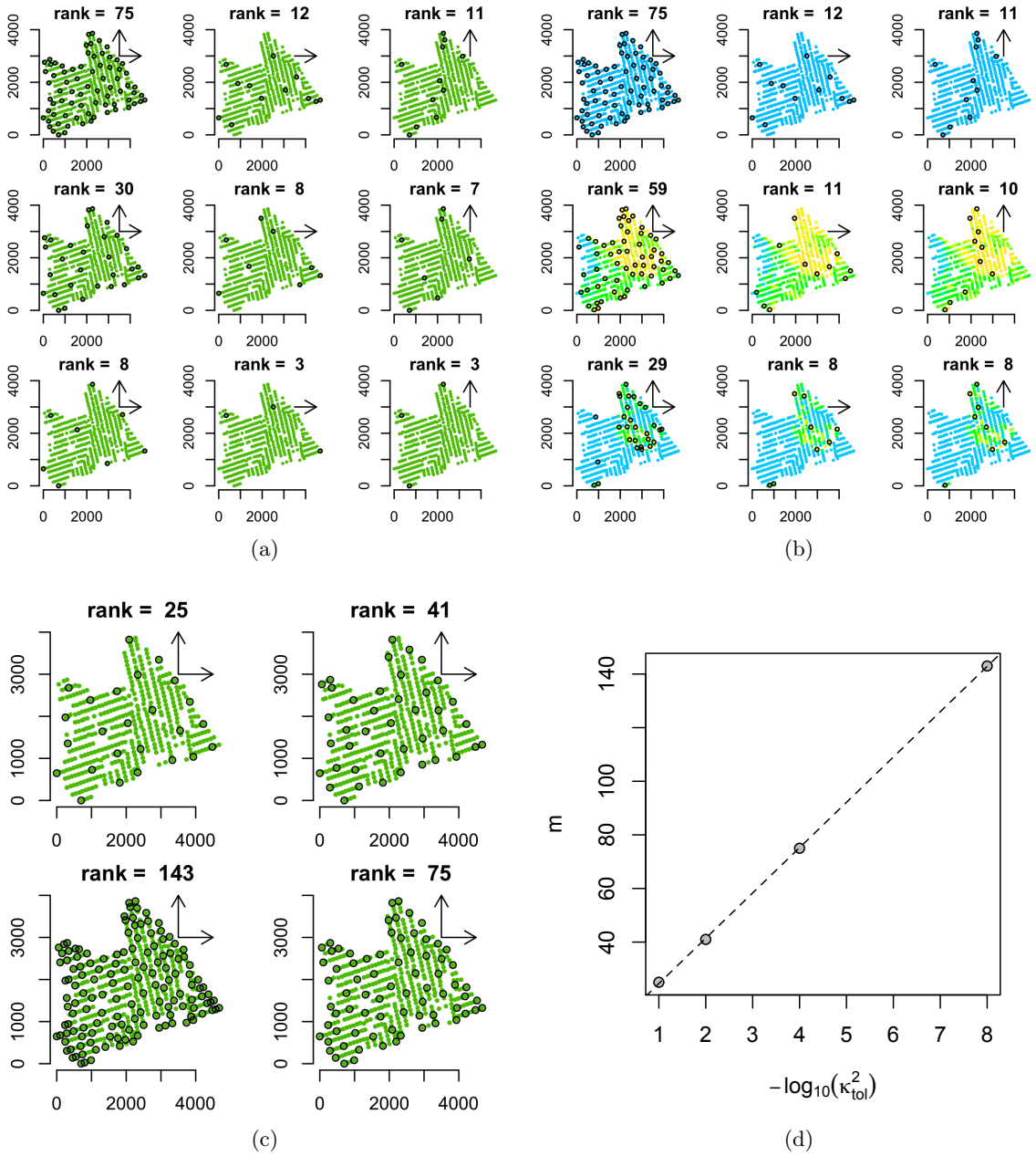


Figure 1: Geometry of knot selection illustrated on Bartlett experimental forest data. (a) Adaptation of knots to changes in Q (columns) and β (rows). (b) The same for changes in Q (columns) and $x(s)$ (rows). (c) The same for changes in κ_{tol}. (d) Number of knots needed to meet specified accuracy bounds for a particular choice of the covariance.

with $1/\kappa^2_{\mathsf{tol}}$. However, theoretical results are not yet available on the relationship between m and κ_{tol}.

4 Application to Bayesian computation

4.1 Sparse nonparametric regression

A low cost, adaptive choice of knots is extremely beneficial in Bayesian Markov chain Monte Carlo (Robert and Casella, 2004; Gilks et al., 1995) computations, where likelihood evaluation is needed at thousands of different values of the covariance parameters. We illustrate this with a sparse non-parametric regression model, where a good exploration of the space of covariance parameters is critical to model fitting. We show that to efficiently explore the covariance parameter space, it is important to adapt to the changes in the canonical metric caused by the changes in these parameter values. In this regard, the adaptive predictive process approximation proposed here has a clear advantage over the existing approaches of handling knots.

We consider a toy data set consisting of (x_i, y_i), $i = 1, \cdots, n = 10,000$, where $x_i = (x_{i1}, \cdots, x_{ip})'$ are drawn independently from the uniform distribution over $[0,1]^p$, with $p = 10$, and y_i are generated independently as $y_i \sim N(2\sin\{2\pi x_{i1}\}, 0.1^2)$. We consider a regression model

$$y_i = \mu + \tau\omega(x_i) + \tau\epsilon_i, \;\; \epsilon_i \overset{\text{IID}}{\sim} N(0, \sigma^2), \tag{7}$$

with ω modeled as a zero-mean Gaussian process over $T = [0,1]^p$. The covariance function of ω is taken to be

$$\psi(s,t) = \psi(s,t|\beta) = \exp\{-\sum_{j=1}^{p} \beta_j^2 (s_j - t_j)^2\}, \;\; t, s \in T.$$

For simplicity of exposition, we set μ and τ^2 at the mean and variance of the observed y_i values, and focus on learning the parameters $\beta_1, \cdots, \beta_p$ and σ^2. The β_j parameters are assigned standard normal prior distributions, folded onto the positive real line and $\log\sigma^2$ is assigned a standard normal prior distribution. Prior independence across parameters is assumed.

A fixed set of knots over T is clearly infeasible for this application due to the dimensionality of T. Placing only 3 knots along each axis takes the total count to $3^{10} = 59049$. While the alternative approach of placing an auxiliary model on the knots and learning them jointly with the covariance parameters via Markov chain sampling can drastically reduce the total number of knots, it is likely to lead to a poor exploration of the covariance parameters for the following reason.

Consider a Gibbs update for β_1 given the remaining parameters and the knots. Suppose the current parameter values are $\beta_1 = \beta_2 = 0.1$, $\beta_3 = \cdots = \beta_p = 0.004$ and $\sigma^2 = \exp(-3.5)$. For these parameter values, a "well learned" choice of the knots is found by applying Algorithm 1 to $\psi(s,t|\beta)$ with β set at the vector of current values (we use $\kappa^2_{\mathsf{tol}} = 10^{-4}$). The corresponding posterior conditional density $p(\beta_1|\beta_2, \cdots, \beta_p, \sigma^2, knots, data)$ is shown by the solid curve in Figure 2. If, instead, the current values had been $\beta_1 = 1$, $\beta_2 = 0.1$, $\beta_3 = \cdots = \beta_p = 0.004$, $\sigma^2 = \exp(-3.5)$ and the knots were chosen by applying Algorithm 1 to the corresponding $\psi(s,t|\beta)$, then the resulting posterior conditional density

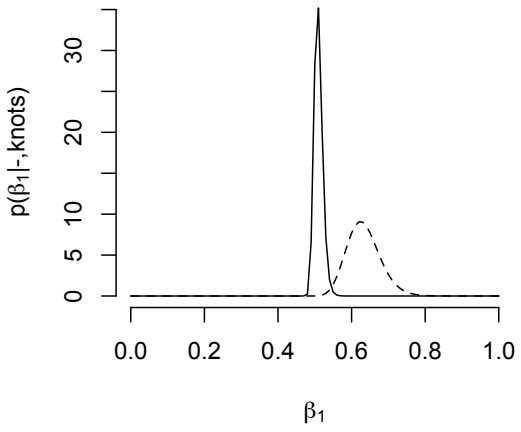


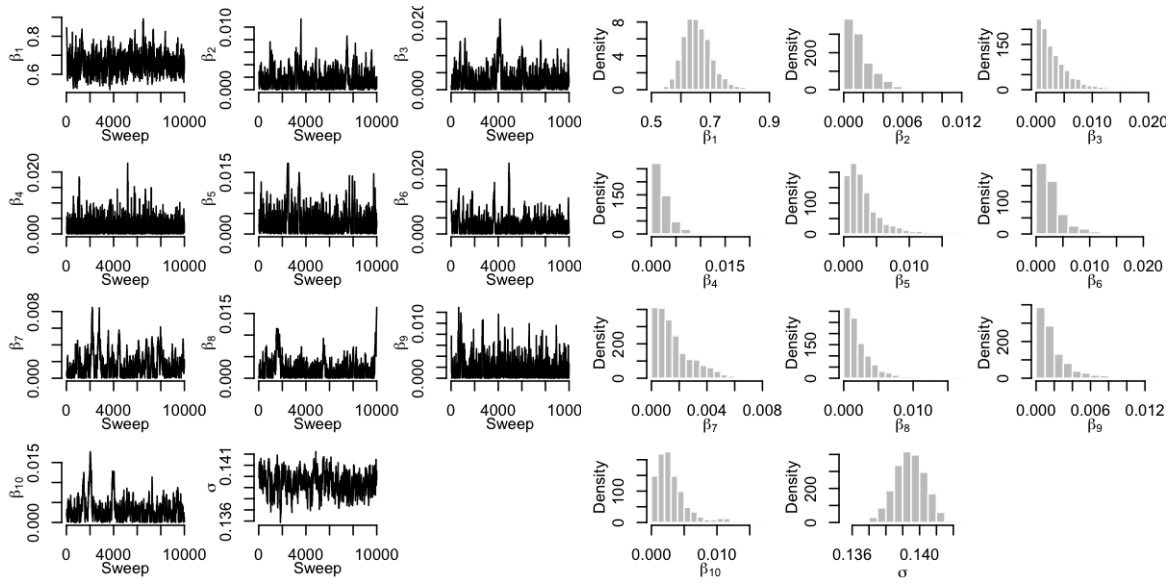
Figure 2: Conditional posterior density of β_1, in the non-parametric regression model, given other parameters and the knots, for two different choices of the knots.

$p(\beta_1|\beta_2, \cdots, \beta_p, \sigma^2, knots, data)$ would look like the dashed curve in Figure 2. The two curves are very different even though the values of the conditioning parameters $\beta_2, \cdots, \beta_p$ and σ^2 are the same. This difference shows that the conditional distribution of the covariance parameters strongly depends on the current set of knots, which, if well learned, should depend on the current values of the covariance parameters. Consequently, there will be additional stickiness in the chain of sampled values of the covariance parameters.

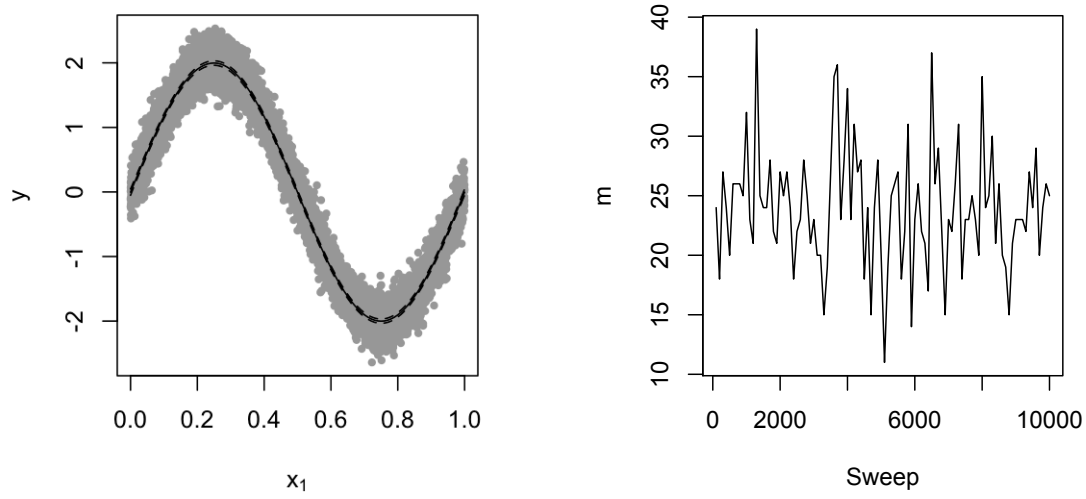
The proposed adaptive predictive process gets rid of this additional stickiness by automatically adapting the knots to the covariance parameter values. As discussed before, for this application it is reasonable to restrict the search of knots to the set of observed covariate vectors $S = \{x_1, \cdots, x_n\}$. Then, one only needs to specify a tolerance level κ_{tol} to obtain an approximation $\hat{p}(y|\beta, \sigma^2)$ of the marginal likelihood $p(y|\beta, \sigma^2) = \int p(y|\omega, \sigma^2)p(d\omega|\beta)$ of (β, σ^2). The approximation obtains by replacing ω in the integral with its predictive process approximation ν adapted to the corresponding $\psi(s, t|\beta)$. The approximate marginal likelihood can be computed in $O(nm^2)$ time, where m is the number of knots needed for $\psi(s, t|\beta)$ with the given tolerance level; detailed formulas are available in Snelson and Ghahramani (2006). The approximate posterior density $\hat{p}(\beta, \sigma^2|data) \propto \hat{p}(y|\beta, \sigma^2)p(\beta, \sigma^2)$ can be explored with common Metropolis-Hastings samplers. Figure 3 reports summaries from a random walk Metropolis sampler exploration with a tolerance level $\kappa^2_{\mathsf{tol}} = 10^{-4}$.

4.2 Varying coefficient regression for spatial data

Pace and Barry (1997) use county level data from 1980 United States presidential election to relate voter turnout to education, income and homeownership standards. They use a spatial autoregressive model to allow this relation vary geographically. We pursue an alternative formulation with a Gaussian process spatial regression model.



(a) (b)



(c) (d)

Figure 3: A random walk Metropolis sampler exploration of the posterior for the non-parametric regression problem and associated posterior summaries. (a) Trace plots of the sampled values of $\beta_1, \cdots, \beta_{10}, \sigma$. (b) Histograms of the same. (c) Posterior median (solid line) and 95% credible intervals (dashed lines) for $f(x)$ at $x = (i/100, 0, \cdots, 0)$, $i = 0, \cdots, 100$, overlaid on the data scatter along x_1 and y. (d) Values of m along the run of the sampler, shown at every 100th sweep.

For county $i = 1, \cdots, n$, let V_i, P_i, E_i, I_i and H_i denote, respectively, voter count, population size of age 18 years or more, population size with at least high school education, aggregate income and number of owner occupied housing units. We take log-percentage voter count $y_i = \log(V_i/P_i)$ as the response variable. Three predictor variables are defined as $x_{1i} = \log(E_i/P_i)$, $x_{2i} = \log(I_i/P_i)$ and $x_{3i} = \log(H_i/P_i)$.

We relate the response to the predictors through a spatially varying regression model

$$y_i = \omega_{0i} + x_{1i}\omega_{1i} + x_{2i}\omega_{2i} + x_{3i}\omega_{3i} + \epsilon_i, \quad \epsilon_i \overset{\text{IID}}{\sim} N(0, \sigma^2). \tag{8}$$

To ensure that geographically proximate counties have similar coefficients, the vector of coefficients from all counties $(\omega_{j1}, \cdots, \omega_{jn})$, for each $j = 0, 1, 2, 3$, is taken to be a zero mean multivariate normal with $\text{Cov}(\omega_{ji}, \omega_{jk}) = \tau_j^2 \exp\{-\beta_j^2 \|t_i - t_k\|^2\}$, where t_i is the spatial location vector of county i, given by the latitude-longitude pair of the county centroid. These four vectors are taken to be mutually independent.

The above formulation is equivalent to

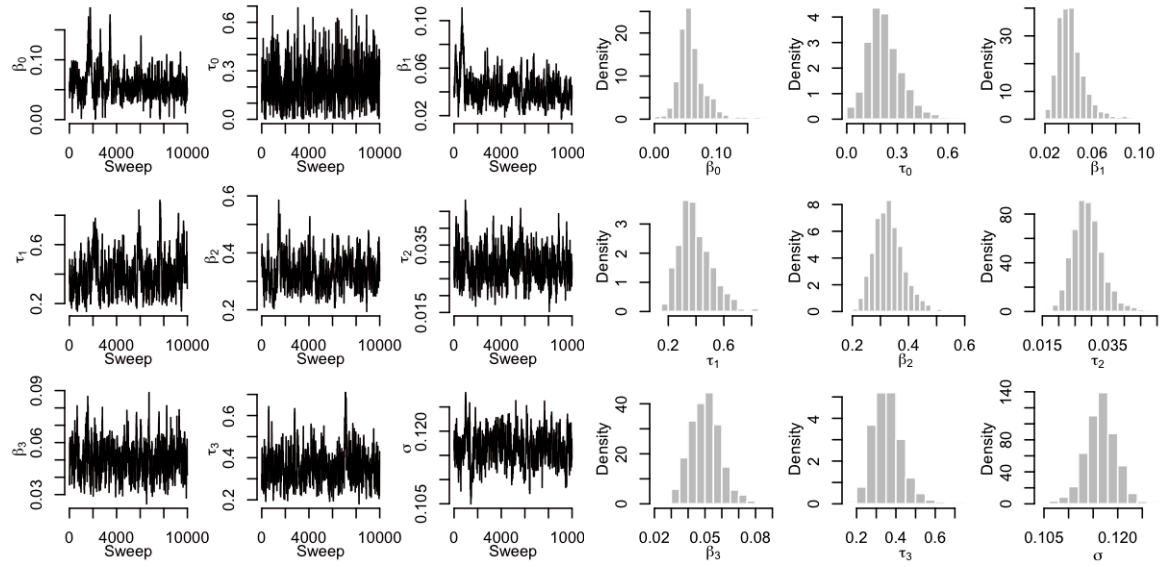
$$y_i = \omega(t_i) + \epsilon_i, \quad \epsilon_i \overset{\text{IID}}{\sim} N(0, \sigma^2) \tag{9}$$

where $\omega(t)$ is a zero-mean Gaussian process on $T = \{t_1, \cdots, t_n\}$ with covariance function $\psi(s, t|\theta) = \sum_{j=0}^{3} x_j(t)x_j(s) \exp\{-\beta_j^2 \|s - t\|^2\}, s, t \in T$, with $x_0(t_i) \equiv 1$, $x_j(t_i) = x_{ji}$, etc, that depends upon the parameter vector $\theta = (\tau_0, \tau_1, \tau_2, \tau_3, \beta_0, \beta_1, \beta_2, \beta_3)$. These covariance parameters are each assigned a standard normal prior distribution, folded onto the positive half of the real line, independently of each other. We also assign $\log \sigma^2$ a standard normal prior distribution, and $\log \sigma^2$ is taken to be a priori independent of θ.

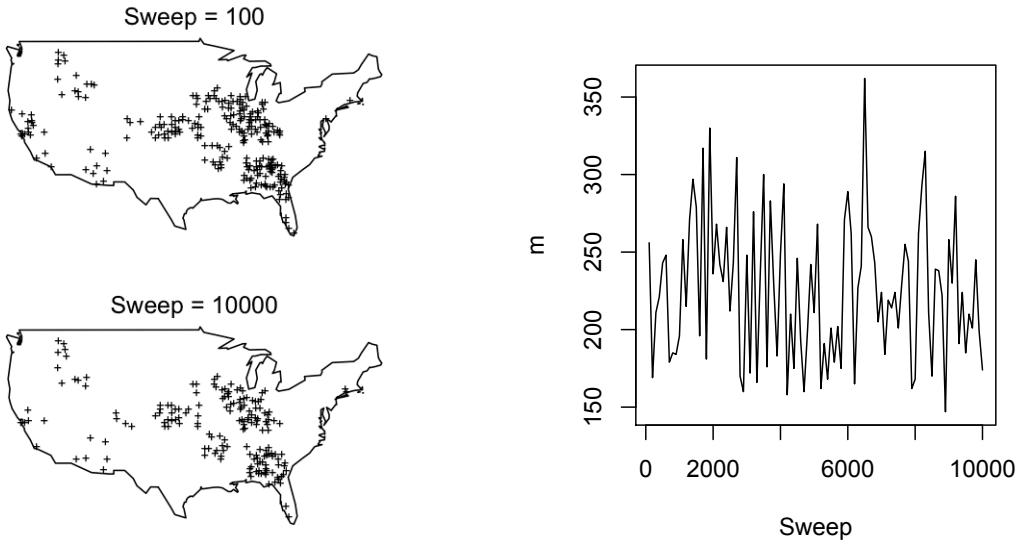
We fit this model using 1040 randomly chosen counties, roughly one third of the total count. Figure 4 shows a random walk Metropolis sampler exploration of the approximate posterior density $\hat{p}(\theta, \sigma^2|y) \propto \hat{p}(y|\theta, \sigma^2)p(\theta)p(\sigma^2)$ with $\kappa_{\text{tol}}^2 = 10^{-4}$ and S taken to be the set of locations for the counties included in model fitting. Trace plots in Figure 4(a) indicate good convergence of the sampler. The marginal posterior distributions of the model parameters appear unimodal (Figure 4(b)). Figure 4(c) shows knots locations from two different sweeps of the sampler. It is interesting that the knots are not uniformly distributed over the whole country. The sampled values of m, shown in Figure 4(d) indicate that unlike the regression example of the previous section a substantial fraction ($\sim$ 15%-35%) of points from S are needed to meet the accuracy bound.

Figure 5 shows summaries of our model fit. Figure 5(a) shows the percentage voter turnout (bottom) and the predicted percentage turnout (top) for all 3106 counties. The predicted value for county i is calculated as the Monte Carlo approximation to $\exp[\text{E}\{\omega(t_i) \mid data\}]$. Figure 5(b) combines the values from Figure 5(a) into a scatter plot, split by counties included in model fitting (green dots) and the remaining ones (red dots). Figure 5(c) shows the spatially varying regression coefficients $\text{E}\{\omega_{ji}|data\}$ on predictors x_{ji}, $j = 1, 2, 3$, for all counties i. Figure 5(d) shows the effect size of these coefficients, found by $\text{E}\{\omega_{ji}|data\}/\sqrt{\text{Var}\{\omega_{ji}|data\}}$.

From Figure 5 we see that predictor x_3, which relates to home ownership, has a strong, spatially varying influence on voter turnout. This predictor has a positive coefficient for all counties, with larger values and effect sizes for counties in the southeast. Predictor x_1, which relates to education, has a moderate, spatially varying influence, with about 1/3 of the counties having an effect size larger than 2. It is interesting that coefficients of x_1 and x_3 appear to be inversely related to each other. Predictor x_2, which relates to income, has

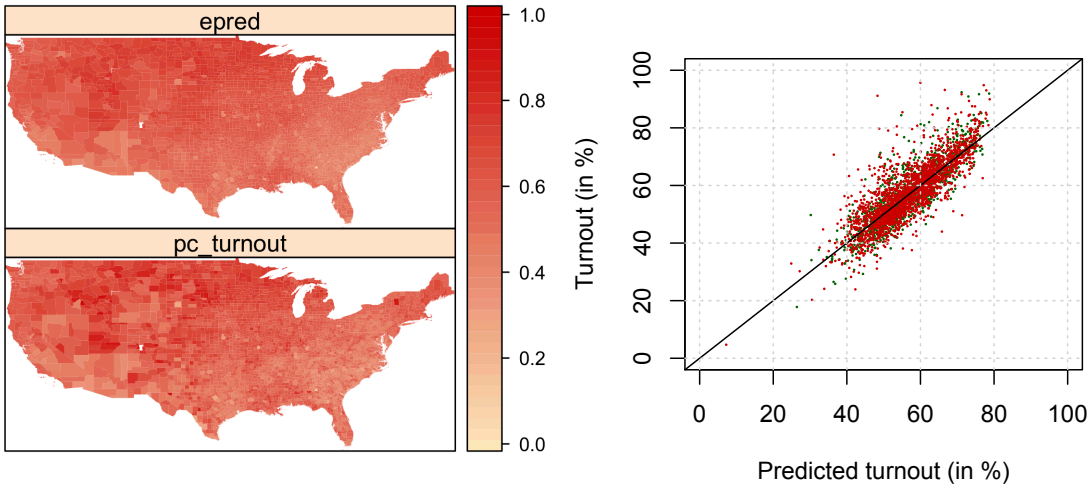


(a) (b)

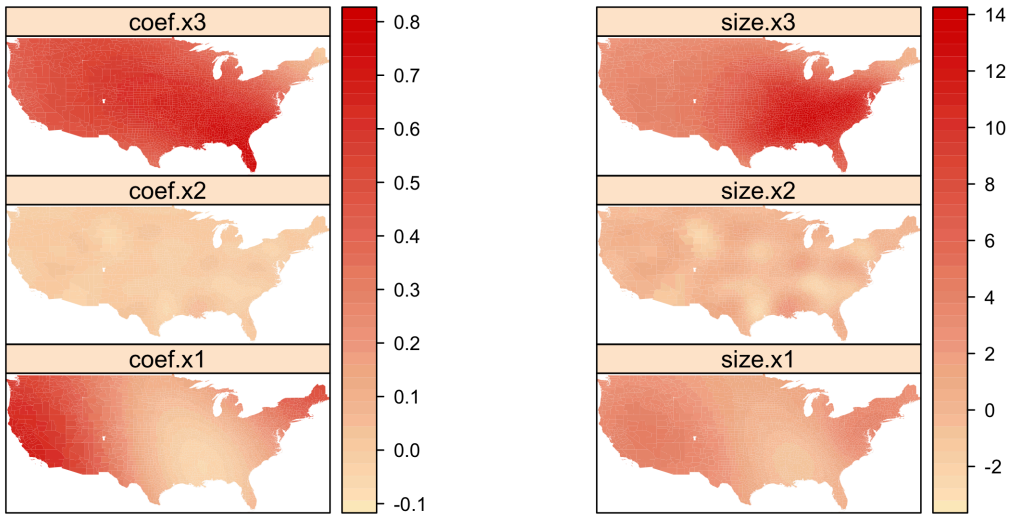


(c) (d)

Figure 4: A random walk Metropolis exploration of the posterior for the election turnout analysis. (a) Trace plots of model parameters. (b) Histograms of the same. (c) Knot locations at two distant sweeps of the sampler. (d) Values of m along the sampler.



(a) (b)



(c) (d)

Figure 5: Posterior summary for the election turnout analysis. (a) Observed (bottom) and predicted (top) percentage turnout shown by counties. (b) Scatter plot of the same, green dots in the background mark training data while the red dots in the foreground are for held out data. (c) Estimated coefficients for the three predictors shown by counties. (d) Effect sizes of the three predictors shown by counties.

a weak influence, with effect size less than 2 in absolute value for more than 95% of the counties. However, the spatial variation of the coefficient of x_2 is more intricate and wavy than that of the other two predictors. Although we do not try to interpret these variations, we note that spatial regression model indeed provides a better fit than an ordinary linear regression that relates y to x_1, x_2 and x_3. The root mean square prediction error from the ordinary linear regression, calculated over the counties not included in model fitting, equals 0.14. The same statistic for the spatial regression model is 0.11.

5 Discussion

We have addressed the question of knot selection within the predictive process methodology and have offered proposals that can substantially automatize the implementation of Gaussian process models in Bayesian analysis. A key conceptual contribution of our work is the emphasis on error bounds to derive a finite rank approximation of an infinite dimensional Gaussian process. It must be noted that the accuracy bounds we provide are all *a priori*. That is, we can not provide an accuracy bound on how well the posterior distribution under the predictive process model approximates the posterior distribution under the original Gaussian process model. However, Tokdar (2007) provides some useful theoretical calculation toward this.

We note that approaches that use an auxiliary model on knots are fundamentally different from our deterministic choice of knots driven by accuracy bounds. Our approach clearly stays within the limits of an approximating method. The smaller the specified tolerance, the closer we are to the original Gaussian process. Approaches with a model on the knots essentially define a different stochastic process. The new stochastic process could indeed provide a better model for the given task, as discussed in Snelson and Ghahramani (2006). However, it would be erroneous to assume that the theoretical properties of a Gaussian process model would also apply to this new stochastic process model.

It is important to realize that at the crux of a predictive process approximation is the rank deficiency of the covariance matrix of the Gaussian vector $W = (\omega(s_1), \cdots, \omega(s_N))$, which is a manifestation of the underlying smoothness of the process ω. For Gaussian process models that employ a relatively un-smooth ω, such as spatial autoregressive models for lattice data (Besag, 1974; Rue et al., 2009), predictive process may not be the ideal approximating tool. Indeed, in such cases, the covariance matrix of W need not be rank deficient, but can have special banded structures that allow for other approximation techniques to yield $O(Nm^2)$ computing.

For a smooth ω, the rank deficiency of the covariance matrix of W poses an additional problem. The covariance matrix, irrespective of its size, can be ill conditioned, making numerical computations unstable. The dynamic, tolerance based stopping of the adaptive predictive process can solve this problem to a large extent, because the knot finding algorithm is likely to terminate before the covariance matrix of $(\omega(s_{\pi_1}), \cdots, \omega(s_{\pi_m}))$ becomes ill conditioned.

A Technical details

Proof of Theorem 2.1. Let $\Psi(x) = \frac{1}{3}e^{x^2}$, $x \geq 0$. $\Psi(x)$ is increasing and convex with $\Psi(0) \in (0, 1)$. For any random variable Z its Ψ-Orlicz norm (Pollard, 1990, page 3) is defined as

$\|Z\|_\Psi := \inf\{C > 0 : \mathrm{E}\Psi(|Z|/C) \le 1\}$. Such a norm provides bounds on tail probabilities as follows: $P(|Z| > x) \le 1/\Psi(x/\|Z\|_\Psi) = 3\exp(-x^2/\|Z\|_\Psi^2)$ for all $x > 0$. This immediately leads to (1) and (2) once we show $\|\sup_{t\in T}|\xi(t)|\|_\Psi \le B\kappa^{1/2}$ and $\|\sup_{t\in S}|\xi(t)\|_\Psi \le 3\kappa_S\sqrt{\log(2+\log|S|)}$.

(i) Lemma 3.4 of Pollard (1990) states that for any $t_0 \in T$,

$$\|\sup_{t\in T}|\xi(t)|\|_\Psi \le \|\xi(t_0)\|_\Psi + \sum_{i=0}^{\infty}\frac{\Delta}{2^i}\sqrt{2+\log D(\frac{\Delta}{2^i},T,d)} \tag{10}$$

$$\le \|\xi(t_0)\|_\Psi + 9\int_0^\Delta \sqrt{\log D(\epsilon,T,d)}d\epsilon \tag{11}$$

where $D(\epsilon,T,d)$ is the ϵ-packing number of T under a (pseudo) metric $d(s,t)$ with $\Delta = \sup_{s,t} d(s,t)$, provided $\|\xi(s)-\xi(t)\|_\Psi \le d(s,t)$ for all $s,t \in T$. It is easy to calculate that $\|Z\|_\Psi = 1.5$ if $Z \sim N(0,1)$ and that $\|Z\|_\Psi = 1.5\sigma$ if $Z \sim N(0,\sigma^2)$. Therefore, for any $s,t \in T$, $\|\xi(s)-\xi(t)\|_\Psi = 1.5[\mathrm{Var}\{\xi(s)-\xi(t)\}]^{1/2}$. Therefore (10) holds if we take $d(s,t) = 1.5[\mathrm{Var}\{\xi(s)-\xi(t)\}]^{1/2}$.

To calculate the right hand side of (10), fix t_0 to be any of the knots used in defining ν. Then $\xi(t_0) = 0$ and consequently the first term $\|\xi(t_0)\|_\Psi = 0$. To calculate the integral in the second term, furst note that

$$d(s,t)^2 = 2.25\mathrm{Var}\{\xi(s)-\xi(t)\} \le 2.25\mathrm{Var}\{\omega(s)-\omega(t)\} \le 2.25c^2\|s-t\|^2$$

where the first inequality holds because $\omega = \nu + \xi$ with ν and ξ independent, and the second inequality follows from our assumption on ω. Therefore $D(\epsilon,T,d) \le \{1 + 1.5c(b-a)/\epsilon\}^d$. Next, bound the diameter Δ as follows

$$\Delta^2 = 2.25\sup_{s,t\in T}\mathrm{Var}\{\xi(s)-\xi(t)\} \le 2.25\sup_{s,t\in T}2[\mathrm{Var}\{\xi(s)\}+\mathrm{Var}\{\xi(t)\}] \le 9\kappa^2. \tag{12}$$

Now use $\log(1+x) \le x$ to bound the right hand side of (10) by

$$9\int_0^{3\kappa}\sqrt{p\log\{1+1.5c(b-a)/\epsilon\}}d\epsilon \le 9\sqrt{1.5pc(b-a)}\int_0^{3\kappa}\epsilon^{-1/2}d\epsilon = B\kappa^{1/2}$$

as desired.

(ii) Now, to calculate $\|\sup_{t\in S}|\xi(t)\|_\Psi$, note that the condition of Lemma 3.4 of Pollard (1990) is trivially satisfied with $d(s,t) = \|\xi(s)-\xi(t)\|_\Psi = 1.5[\mathrm{Var}\{\xi(s)-\xi(t)\}]^{1/2}$ due to discreteness of S and $D(\epsilon,T,d) \le |S|$ for all $\epsilon > 0$. Therefore we can apply (10) with S instead of T to conclude $\|\sup_{t\in S}|\xi(t)|\|_\Psi \le \Delta\sqrt{2+\log|S|}$ where $\Delta^2 = \sup_{s,t\in S}d(s,t)^2 \le 9\kappa_S^2$ as in (12).

□

References

Adler, R. J. (1990). *An introduction to continuity, extrema, and related topics for general Gaussian processes*, Volume 12. Hayward, CA: Institute of Mathematical Statistics.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society Series B 70*(4), 825–848.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B 36*, 192–209.

Choi, T. and M. Schervish (2007). On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis 98*(10), 1969–1987.

Csató, L., E. Fokoué, M. Opper, B. Schottky, and O. Winther (2000). Efficient approaches to gaussian process classification. In S. A. Solla, T. K. Leen, and K.-R. Müller (Eds.), *Advances in Neural Information Processing Systems*, Volume 12, Cambridge, MA. The MIT Press.

Finley, A. O., H. Sang, S. Banerjee, and A. E. Gelfand (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis 53*(8), 2873–2884.

Foster, L., A. Waagen, N. Aijaz, M. Hurley, A. Luis, J. Rinsky, C. Satyavolu, M. J. Way, P. Gazis, and A. Srivastava (2009). Stable and efficient gaussian process calculations. *The Journal of Machine Learning Research 10*, 857–882.

Ghosal, S. and A. Roy (2006). Posterior consistency of gaussian process prior for nonparametric binary regression. *The Annals of Statistics 34*, 2413–2429.

Gilks, W., S. Richardson, and D. Spiegelhalter (1995). *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*. Chapman and Hall/CRC.

Guhaniyogi, R., A. O. Finley, S. Banerjee, and A. E. Gelfand (2010). Adaptive gaussian predictive process model for large spatial data sets. American Statistical Association Joint Statistical Meeting. August 2, 2010. Vancouver, British Columbia.

Handcock, M. S. and M. L. Stein (1993). A bayesian analysis of kriging. *Technometrics 35*, 403–410.

Kennedy, M. and A. O'Hagan (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society, Series b 63*, 425–64.

Kim, H.-M., B. K. Mallick, and C. C. Holmes (2005). Analyzing nonstationary spatial data using piecewise gaussian processes. *Journal of the American Statistical Association 100*, 653–668.

Neal, R. M. (1998). Regression and classification using gaussian process priors. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics*, Volume 6, pp. 475–501. Oxford University Press.

Oakley, J. and A. OHagan (2002). Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika 89*, 769–784.

Pace, R. K. and R. Barry (1997). Quick computation of spatial autoregressive estimators. *Geographical Analysis 29*, 232–247.

Petrone, S., M. Guindani, and A. E. Gelfand (2009). Hybrid dirichlet mixture models for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(4), 755–782.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, Volume 2. Institute of Mathematical Statistics and American Statistical Association.

Quiñonero-Candela, J. and C. E. Rasmussen (2005). A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research 6*, 1939–1959.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods* (2 ed.). Springer.

Rue, H., S. Martino, and N. Chopin (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society, Series B 71*, 319–392.

Schwaighofer, A. and V. Tresp (2003). Transductive and inductive methods for approximate gaussian process regression. In S. Becker, S. Thrun, , and K. Obermayer (Eds.), *Advances in Neural Information Processing Systems*, Volume 15, Cambridge, Massachussetts, pp. 953–960. The MIT Press.

Seeger, M., C. K. I. Williams, and N. Lawrence (2003). Fast forward selection to speed up sparse gaussian process regression. In C. M. Bishop and B. J. Frey (Eds.), *Ninth International Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics.

Shi, J. Q. and B. Wang (2008). Curve prediction and clustering with mixtures of gaussian process functional regression models. *Statistical Computing 18*, 267–283.

Short, M. B., D. M. Higdon, and P. P. Kronberg (2007). Estimation of faraday rotation measures of the near galactic sky using gaussian process models. *Bayesian Analysis 2*, 665–680.

Smola, A. J. and P. L. Bartlett (2001). Sparse greedy gaussian process regression. In *Advances in Neural Information Processing Systems*, Volume 13, Cambridge, Massachussetts, pp. 619–625. The MIT Press.

Snelson, E. and Z. Ghahramani (2006). Sparse gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. Platt (Eds.), *Advances in Neural Information Processing Systems*, Volume 18, Cambrisge, Massachussetts. The MIT Press.

Sudderth, E. and M. Jordan (2009). Shared segmentation of natural scenes using dependent pitman-yor processes. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21*. MIT Press.

Tokdar, S. T. (2007). Towards a faster implementation of density estimation with logistic gaussian process priors. *Journal of Computational and Graphical Statistics 16*, 633–655.

Tokdar, S. T. and J. K. Ghosh (2007). Posterior consistency of logistic gaussian process priors in density estimation. *Journal of Statistical Planning and Inference 137*, 34–42.

Tokdar, S. T. and J. B. Kadane (2011). Simultaneous linear quantile regression: A semiparametric bayesian approach. Duke Statistical Science Discussion Paper #12.

Tokdar, S. T., Y. M. Zhu, and J. K. Ghosh (2010). Density regression with logistic gaussian process priors and subspace projection. *Bayesian Analayis 5*(2), 316–344.

van der Vaart, A. W. and J. H. van Zanten (2008). Rates of contraction of posterior distributions based on gaussian process priors. *Annals of Statistics 36*, 1435–1463.

van der Vaart, A. W. and J. H. van Zanten (2009). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annal of Statistics 37*(5B), 2655–2675.

Zhu, Z. and M. Stein (2005). Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference 134*, 583–603.

Zimmerman, D. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics 17*, 635–652.